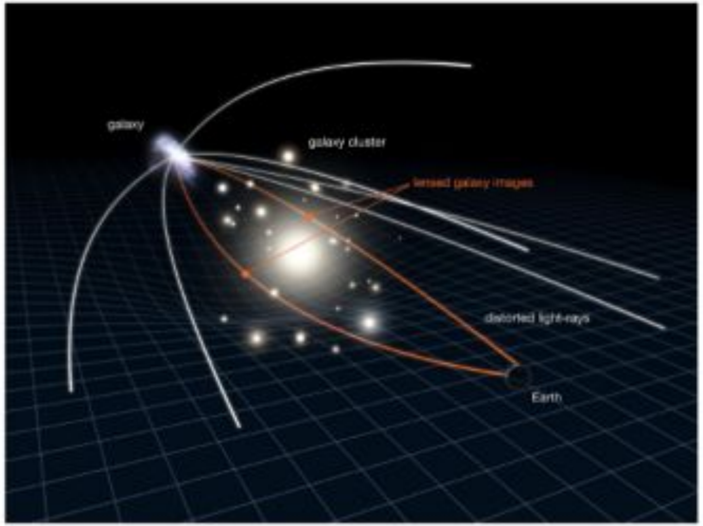
Machine Learning for Astrophysics : Identifying Blended Sources in Galaxy Images

Alexandre BRUCKERT
Supervisor : Samuel FARRENS

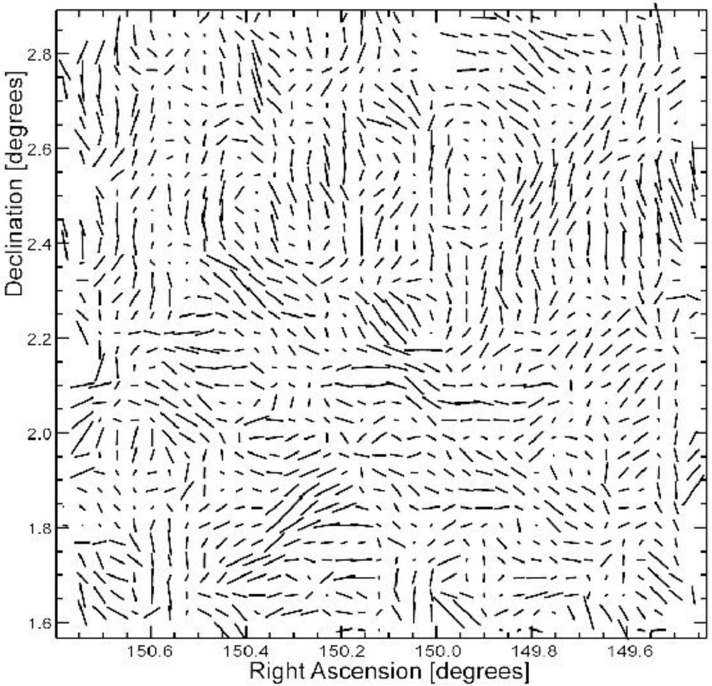
Dark matter and weak lensing

- Dark matter : theoretical matter, that would represent around 85% of the total mass of the universe
- Only reacts to gravitational forces
- What is dark matter, and what is its distribution ?
- Weak lensing : (very) small shear in the observed galaxies because of huge foreground masses
- Statistical methods to compute that shear, and find mass maps



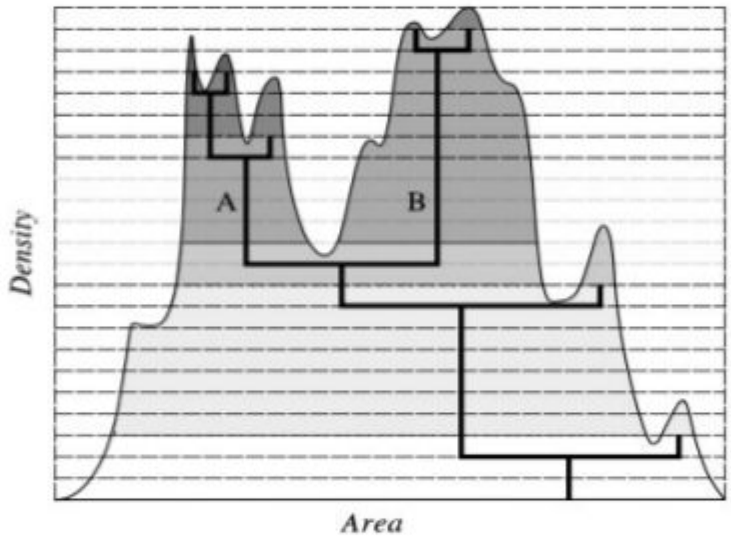
Deblending

- Overlapping of two or more sources in an image
- Many different reasons : line of sight, PSF, shear...
- Very different rates of occurence
- Issue when it comes to compute the shear
- Two solutions : get rid of those objects, or separate them
- Several problems to solve : identification (binary classification), sources count (multi-class or regression), finding the contours of each objects (segmentation)...



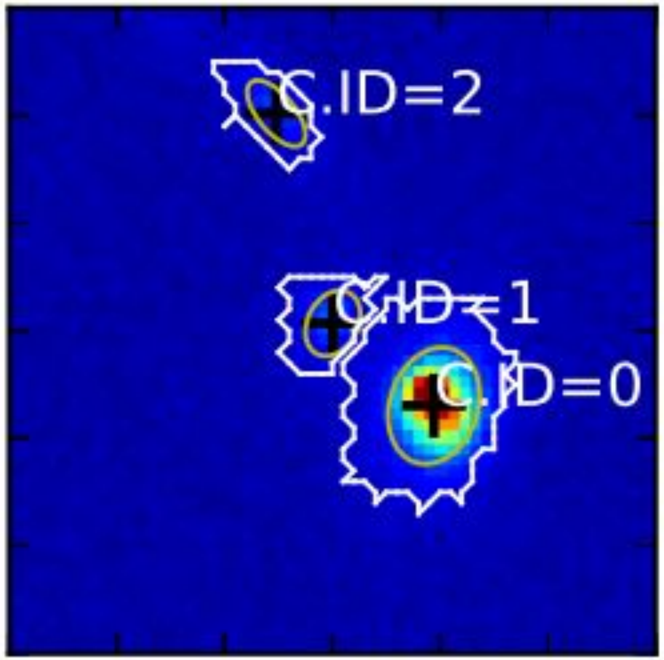
SExtractor

- Most used tool to detect and extract light sources
- Includes a deblending module, that identifies blended sources
- Threshold-based method
- Problems : relies on human hand to set the thresholds, and is not very accurate



Other methods

- Several different techniques, mostly for specific surveys
- ASTErIsM : clustering based
- PCA-based methods
- Flux measurements
- No universal method that works well in most of the cases

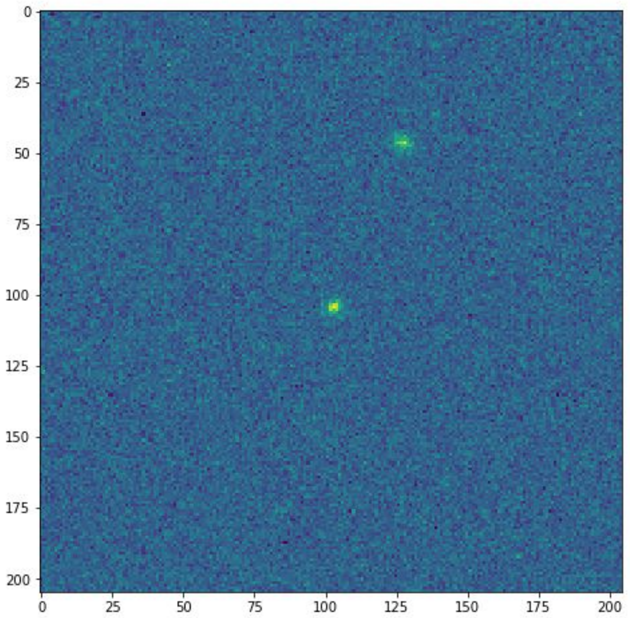


Datasets and simulations

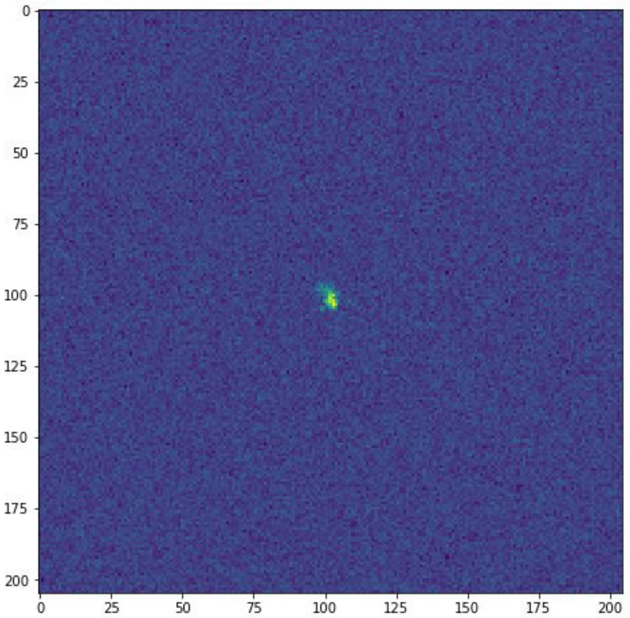
- Three simulated datasets, using GalSim to be generated
- GREAT3 : 20 000 images (10 000 of each class), basic simulations
- Euclid : 10 000 images (7500 blends, 2500 non-blends), based on expected Euclid images
- CFIS : unlimited amount of images (usually, 40 000 when it comes to run the model), high-quality simulations based on the CFIS survey
- Three different kind of images : blends of two sources, single sources and two separated sources.
- Simulation of blends : generate a first galaxy at the center, and randomly place another one on the image, until the two of them are blended

Examples of simulations - GREAT3

Non-blended

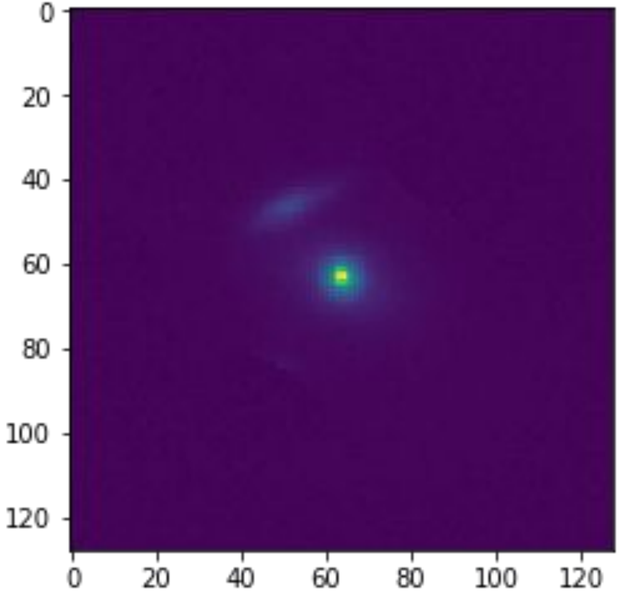


Blended

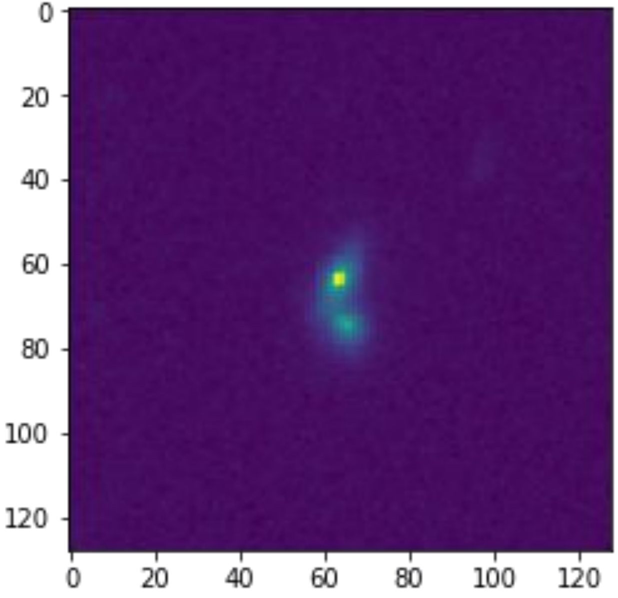


Examples of simulations - Euclid

Non-blended

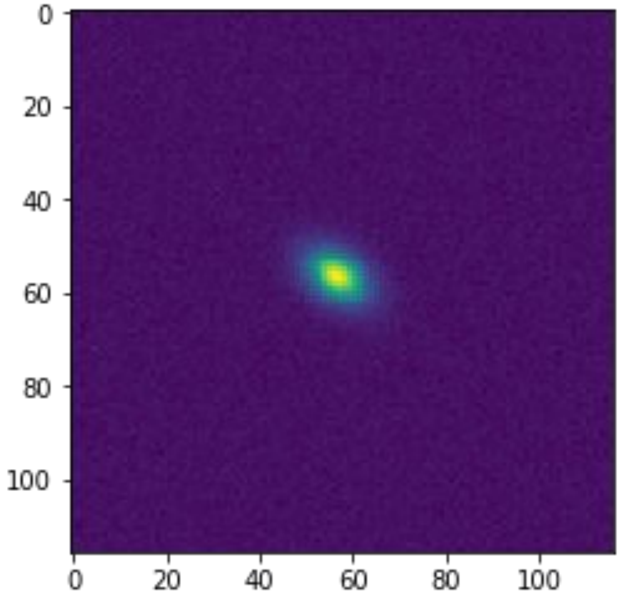


Blended

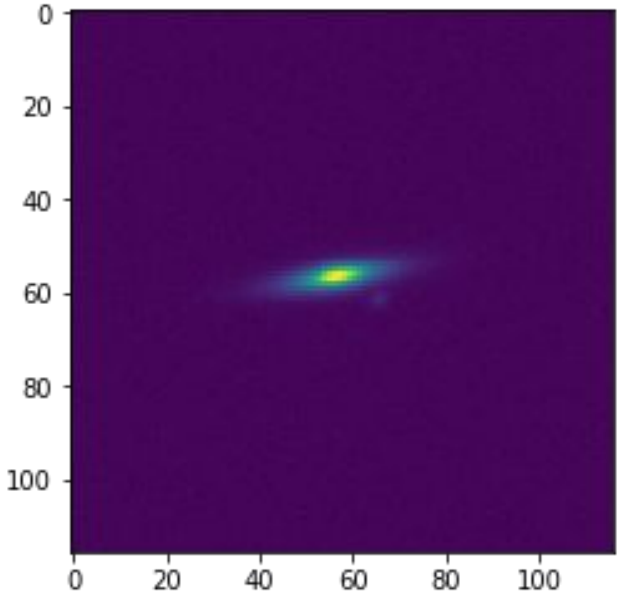


Examples of simulations – CFIS

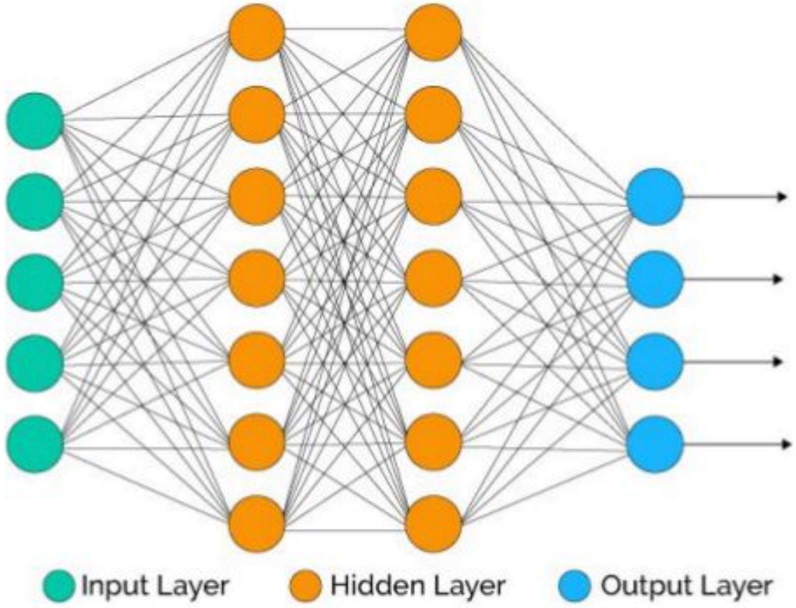
Non-blended

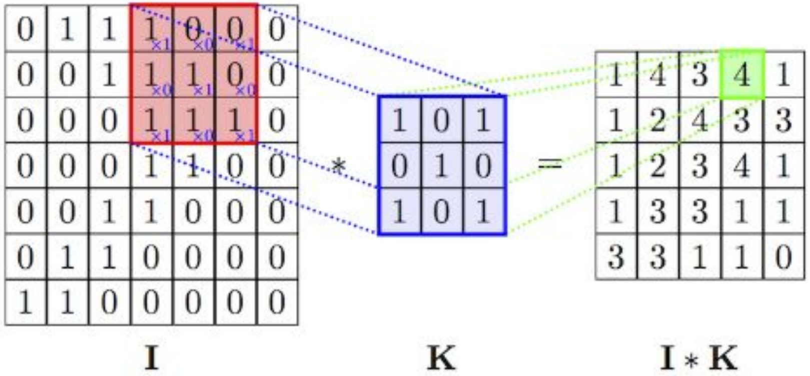


Blended

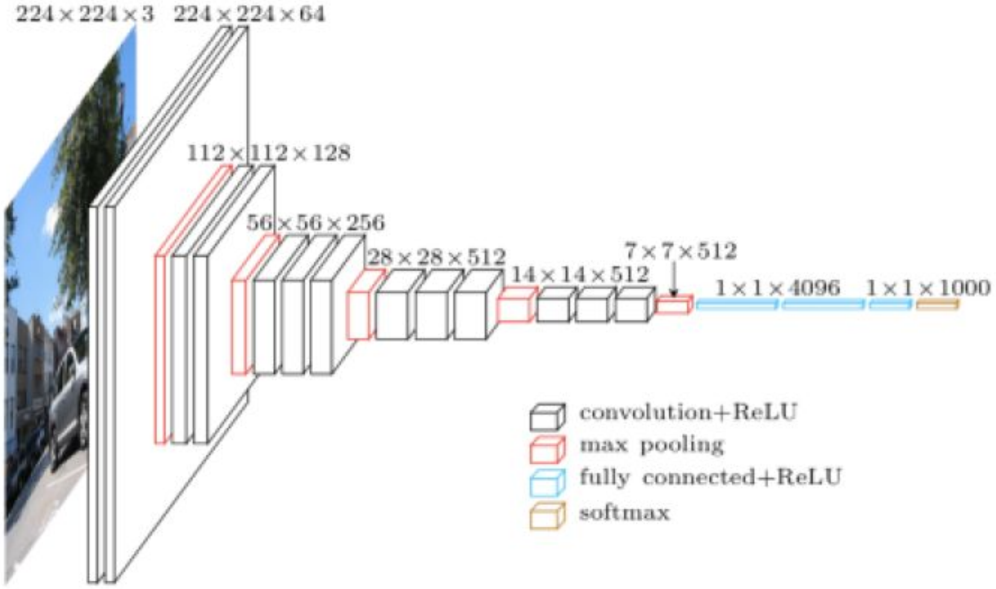


Quick overview of CNNs





VGG16 image classifier



- Pre-trained network (time of training reduced)
- Popular and effective network for classification
- Very simple architecture
- No shortcut, normalization or concatenation operations

Very Deep Convolutional Network For Large-scale Image Recognition, K. Simonyan & A. Zisserman

Results – Methods comparison – GREAT3

VGG16		Actual labels	
		Blended	Non-blended
Predicted labels	Blended	0,963	0,023
	Non-blended	0,037	0,977

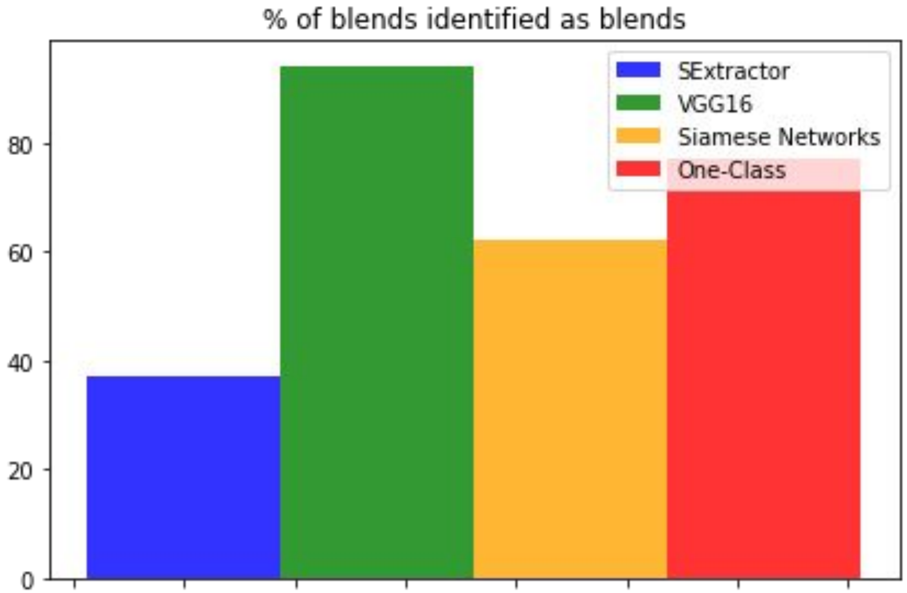
One-class		Actual labels	
		Blended	Non-blended
Predicted labels	Blended	0,794	0,139
	Non-blended	0,206	0,861

Siamese networks		Actual labels	
		Blended	Non-blended
Predicted labels	Blended	0,672	0,144
	Non-blended	0,328	0,856

SExtractor		Actual labels	
		Blended	Non-blended
Predicted labels	Blended	0,565	0,245
	Non-blended	0,435	0,755

Results - Methods comparisons - GREAT3

When the noise get higher (σ > 5e-3), the problems of SExtractor appear even more



Results – CFIS (trained on GREAT3)

VGG16		Actual labels	
		Blended	Non-blended
Predicted labels	Blended	0,824	0,110
	Non-blended	0,176	0,890

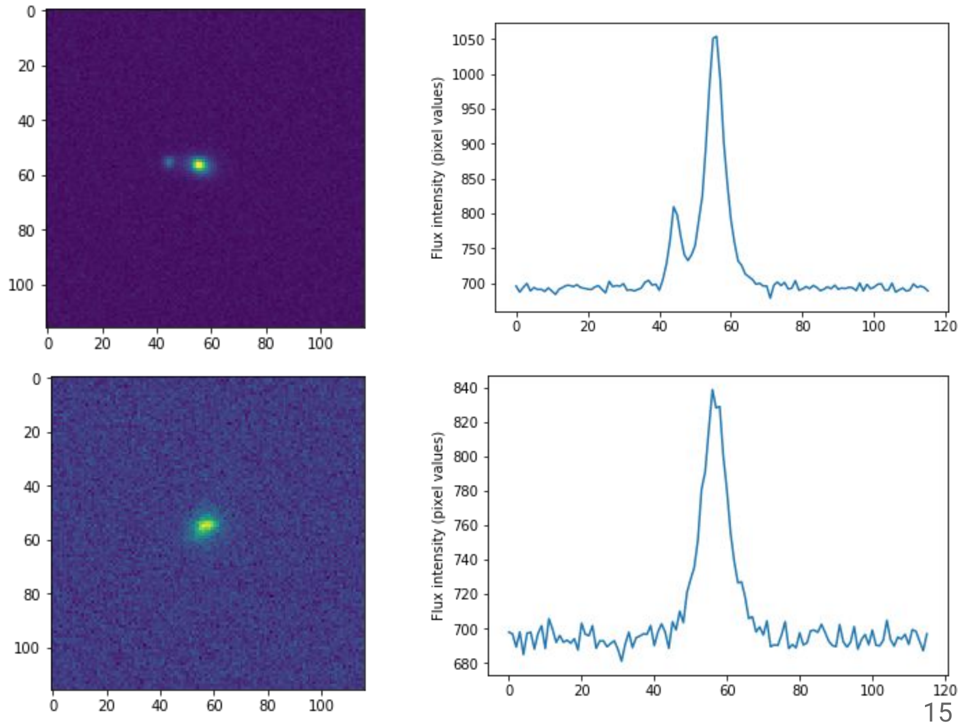
One-class		Actual labels	
		Blended	Non-blended
Predicted labels	Blended	0,617	0,164
	Non-blended	0,383	0,836

Siamese networks		Actual labels	
		Blended	Non-blended
Predicted labels	Blended	0,552	0,282
	Non-blended	0,448	0,718

SExtractor		Actual labels	
		Blended	Non-blended
Predicted labels	Blended	0,453	0,131
	Non-blended	0,547	0,869

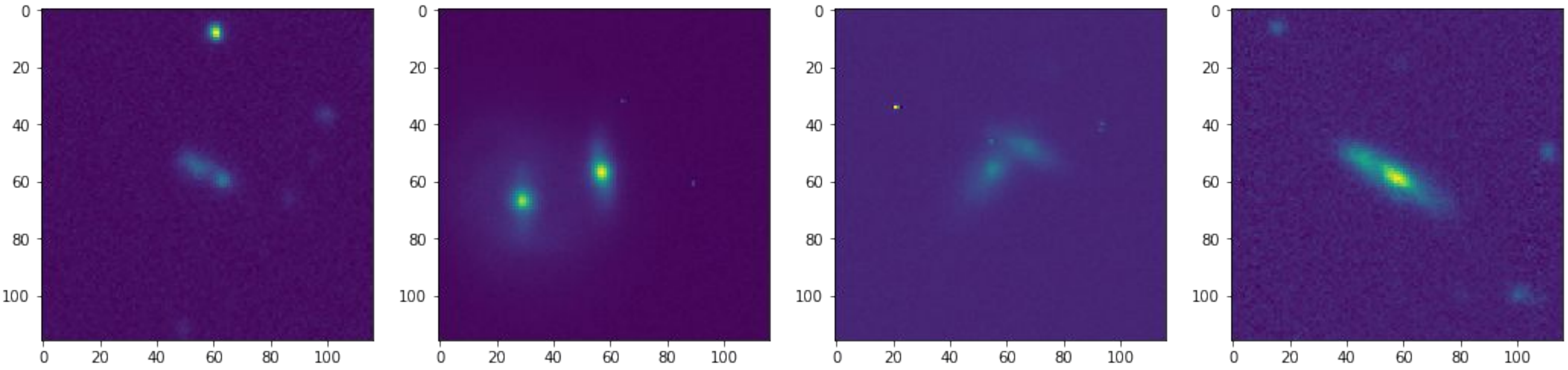
VGG16 – Results analysis

- Very good general accuracy (95,48% on CFIS, when trained on a mixed dataset)
- Reasons of misidentifications :
  - Sources too close to each other
  - Lower signal-to-noise ratio
  - Discrepancies in light intensity



A few real CFIS-results

Obvious blended objects properly identified



Current Work

- Creating a database of blended sources from real images to check whether or not the network performs well on more realistic images
- Running the network on real CFIS images and analysing the flag differences between the different methods
- Running shape measurement in several situations :
    - With all the sources
    - Removing the blended sources found by SExtractor
    - Removing the blended sources found by VGG16

Future work

- Extending the model to multi-class classification, in order to count the number of sources
- Improving the simulations techniques to be closer to real data (using GANs, for instance)
- Applying segmentation techniques to detect overlapping zones, and create masks for the actual deblending (SSD, Mask-RCNN, …)

Thank you for your attention !